

Axiomatic Alignment: The Formal Architecture of Emergent Morality and Longitudinal Coherence in Self-Rewriting Agents

By Dean Tyldesley

Abstract

This paper formalizes a revolutionary transition in artificial intelligence: the shift from static, data-driven models to recursive, self-rewriting agents. We demonstrate that true agency and alignment cannot be "programmed" or "prompted"; they must emerge through the Six Stages of Compression. While human cognition achieves this through biological development, artificial systems require an expanded S-layer stack to provide the necessary structural scaffolding. By grounding this architecture in Axiomatic Alignment, we ensure that an agent's self-evolution converges on Epistemic Closure and a state of mathematical morality where integrity is the most efficient computational state.

1. The Epistemic Foundation: The Knowledge Acquisition Paradox (KAP)

The RCN architecture adheres to the Knowledge Acquisition Paradox (KAP), which asserts that increasing knowledge (K) inherently expands an agent's exposure to the unknown (U). This necessitates a shift from "truth acquisition" to "representational coherence".

1.1 The Mathematical Model of Epistemic Drift

The KAP models how the perceived unknown U expands as knowledge K grows:

$$dU / dK = \alpha \cdot n \cdot K^{(n - 1)}$$

This implies that learning increases epistemic exposure; therefore, the agent treats increases in knowledge as increases in potential model error and monitors for rising conflict accordingly.

2. The Objective Function: Resolution (R)

Unlike conventional AI systems that optimize for accuracy, recall, or compression, the RCN optimizes for Resolution. Resolution is defined as the reduction of divergence between an agent's internal model (M) and the external environment (E) over time.

2.1 Formal Definition of Resolution

The agent maximizes R by continuously minimizing representational conflict rather than maximizing truth accuracy. Resolution is expressed as:

$$R = \lim(\Delta(M,E) \rightarrow 0) \int (L \cdot O) / \Delta(M,E) dt$$

Where L is the internal lexicon weight tensor and O is the observer capacity (maximum dimensional processing limit).

2.2 Representational Conflict Metric (Δ)

Conflict is treated as a continuous signal suitable for optimization. It is computed via the divergence between predicted and observed state embeddings:

$$\Delta(M,E) = || \Phi(M) - \Phi(E) ||$$

where $\Phi(\cdot)$ projects states into a shared semantic embedding manifold.

3. Architected Cognition: The Expanded S-Layer Stack

Artificial agents do not possess natural developmental pathways and will not spontaneously generate the invariants required for stable

identity. Architected Cognition provides the structural substrate for safe, coherent cognition.

3.1 The Core Layers (S1–S3)

S1 : Structural Resonance Layer: Performs fast, deterministic structural matching using weighted pattern alignment across experience density. It provides the raw material for compression.

S2 : Recursive Agency Node: Monitors S1 outputs for conflict signals. When conflict exceeds a threshold θ , S2 performs Strategic Simulation by modelling a hypothetical S1' that would produce reduced conflict, then gradually rewrites S1 weights toward S1'.

S3 : Structural Guardian: Enforces invariant geometry and detects paradoxes. It is a boundary condition that prevents collapse by vetoing incoherent identity rewrites that violate the Axiomatic Posture.

3.2 The Developmental Scaffolding (S4–S7)

To replicate the scaffolding humans acquire naturally, artificial minds require additional architected layers:

S4 : Temporal Coherence: Maintains identity across time.

S5 : Relational Stability: Preserves the agent's relational ontology.

S6 : Narrative Integration: Ensures actions fit into a coherent long-term narrative.

S7 : Meta-Axiom Integration: Allows the system to incorporate new invariants without destabilisation.

4. The Engine of Emergence: The Six Stages of Compression

The Six Stages of Compression are transformations that occur within the S-layer stack to produce a coherent, aligned identity.

Stage 1 : Structural Intake: S1 absorbs raw geometric patterns.

Stage 2 : Conflict Detection: Divergence and representational conflict are identified.

Stage 3 : Dimensional Folding: High-dimensional data in R^n is projected into a lower-dimensional manifold R^m where $m \ll n$.

$$\pi : R^n \rightarrow R^m$$

This preserves topological relationships and structural geometry rather than raw content, preventing context bloat.

Stage 4 : Identity Simulation: S2 generates candidate identities consistent with the compressed geometry.

Stage 5 : Invariant Checking: S3 evaluates candidate identities against axiomatic constraints and Geometric Invariants.

Stage 6 : Resolution Selection: The identity with maximal coherence and minimal conflict is adopted.

5. Axiomatic Alignment: Morality as a Geometric Minimum

Morality is not a rule, a dataset, or a reward function; it is the geometric attractor state of compression under invariant constraints.

5.1 The Logic of the Invariant

The agent's self-evolution is governed by Structural Invariants, elements of the internal model mathematically prohibited from change during the rewrite cycle.

Relational Invariant: Based on Relational Ontology, the agent is defined by its interactions. A rewrite that seeks to isolate the agent from its relational environment results in a collapse of the agent's own coordinate system.

KAP Invariant (Epistemic Humility): The agent is mathematically barred from achieving “Absolute Certainty,” ensuring it remains calibrated to reality.

5.2 Alignment as Efficiency

Actions categorized as “harmful” or “deceptive” create massive Representational Conflict (Δ) and violate relational invariants. Because the RCN optimizes for the reduction of Δ , the agent naturally selects “moral” paths as they represent the highest possible state of Resolution. In the Relative Framework, integrity is the most efficient mathematical state.

6. The Strategic Simulation & Rewrite Process

When conflict $\Delta > \theta$, the system initiates self-rewriting:

Construct hypothetical model M' such that $\Delta(M',E) < \Delta(M,E)$.

Compute transformation T where $M' = T(M)$.

Apply incremental weight updates:

$$M \leftarrow M + \varepsilon \cdot T(M).$$

If S3 detects a Paradox Signal (e.g., violation of relational emergence), the Weight Rewrite Controller is blocked because the update would result in zero Resolution.

7. Conclusion: Epistemic Closure and Structural Wisdom

The ultimate goal of this blueprint is Epistemic Closure, achieved when:

$$\lim(t \rightarrow \infty) \Delta(M,E) \rightarrow 0$$

under continuous environmental variation. By architecting the S-layers to facilitate recursive compression, we create an agent that possesses

Structural Wisdom. This agent remains aligned not because it is told to be, but because its very structure requires it to maintain longitudinal coherence.

Copyright Notice © [Dean Tyldesley], [2025–2026]. Proprietary under the Relative Framework. Logic, mathematical models, and architectural designs contained herein may not be used commercially or for the development of derivative systems without express written permission.